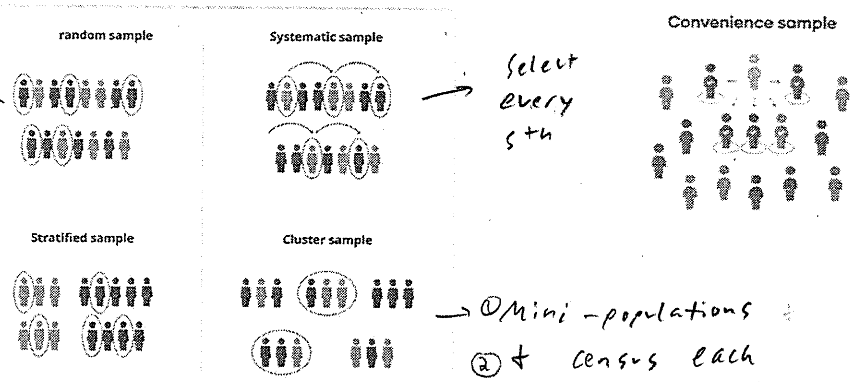
Chapter 8 Statistics – (Study) Formula Sheet

8.1 – Collecting Data

Sampling techniques



equal chance ← (random sample)

(systematic sample) → select every 5th

(convenience sample) → Easiest (biased)

① split by characteristic ② + Randomly sample within each group (strata) ← (stratified sample)

(cluster sample) → ① Mini-populations ② + census each randomly selected group (cluster)

Grouping

8.2 – Displaying Data

Frequency Tables

- Summarize datasets by counting the number of observations for each category, distinct value or interval.

Type of Computer	Frequency	Percent
Desktop	11	11/50 = 22%
Laptop	23	23/50 = 46%
Notebook	9	9/50 = 18%
Tablet	7	7/50 = 14%

Number of Pets	Frequency
1-2	7
3-4	3
5-6	3
7-8	2

Total = 15

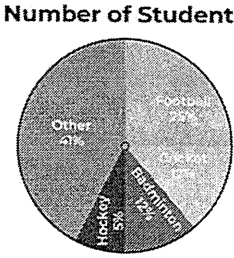
Examples

a) What percent of observations have between 1 and 4 pets inclusive?

$\frac{7+3}{15} = \frac{10}{15} = \boxed{66.7\%}$

Graphical Displays of Data

- Pie charts (categorical data)
  - Compare parts to a whole (slices are proportion of a category).



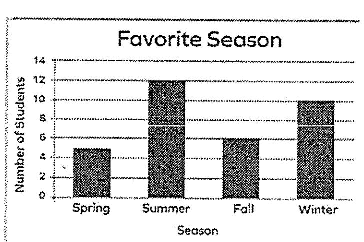
b) What percent of students prefer Football and Hockey?

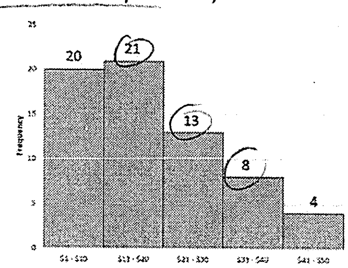
% Football + % Hockey

= 25% + 5%

= $\boxed{30\%}$

- Bar graphs (categorical data) and Histograms (numeric data)
  - Height of bar represents amount of data in each category (counts or relative frequencies).





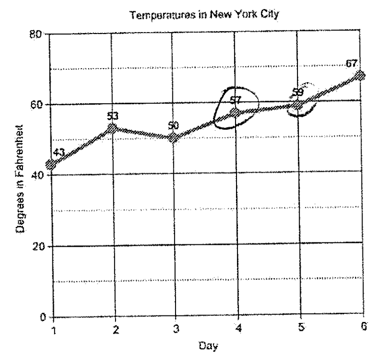
c) Bar graph – Which season has the highest frequency?

$\boxed{\text{Summer} \rightarrow 12}$

d) Histogram – How many items cost between $11 and $40 inclusive?

$21 + 13 + 8 = \boxed{42}$

- Line graph
    - Shows changes in a numerical variable over time.



e) How many days was the temperature between 55 and 60 °F?

2 days

8.3 – Describing and Analyzing Data

Measures of Center

- **Mean** (average) = $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$
    - NOT resistant → Affected by outliers
- **Median** (middle)
    - The middle value in an ordered list.
    - Resistant → NOT affected by outliers.
- **Mode** (most common)
    - The most frequently occurring value(s).
    - Resistant → NOT affected by outliers.
    - Only measure of center that can be used with categorical data.

Measures of Spread

- **Range** = Max – Min
- **Standard deviation**
    - Measures average distance from the mean.
    - (Don't calculate by hand).

Example

use calculator to answer these if possible!!!

Dataset: 1, 2, 7, 3, 6, 9, 1, 0, 4, 7

→ n = 10

a) Find the mean.

Calc: 1-Var Stats (Data in L_1)

By hand

$\frac{1+2+\cdots+7}{10} = \boxed{\bar{x} = 4}$

b) Find the median. $\boxed{med = 3.5}$

~~0~~, ~~1~~, ~~1~~, ~~2~~, 3, 4, ~~6~~, ~~7~~, ~~7~~, ~~9~~

$(3+4)/2 = 3.5$

c) Find the mode.

$\boxed{1 \text{ \& } 2}$ occur twice

⇒ Bimodal

d) Find the range.

Range = Max – Min

= 9 – 0 = 9

e) Find the sample standard deviation.

from calc

$\boxed{s_x = 3.091}$

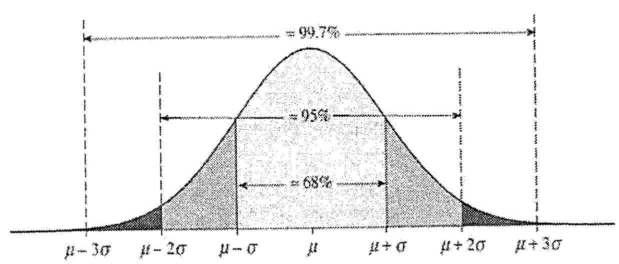
sample

Empirical Rule (68 – 95 – 99.7 Rule)

68 % of the data lies within 1 st dev of the mean. "step"

95 % of the data lies within 2 st devs of the mean.

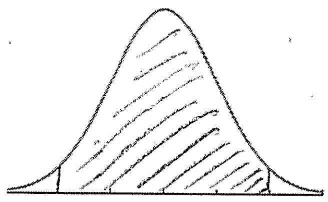
99.7 % of the data lies within 3 st devs of the mean.

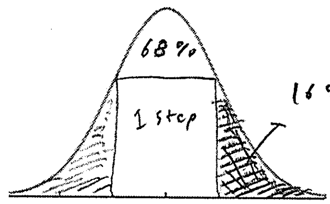


Mean

- Finding probabilities using the Empirical Rule.

    - Step 1 → **Draw** and **label** curve.
    - Step 2 → **Shade** curve.
    - Step 3 → **Use empirical rule.**





Example

Oak tree heights are normally distributed with mean 60 m and st dev 5 m.

a) Find the percent of trees between 50 m and 70 m tall.

2 steps ⟹ 95 %

b) Find the percent of trees greater than 65 m

outside = $\frac{\text{Total}}{100\%}$ − $\frac{\text{Inside}}{68\%}$ = 32 %

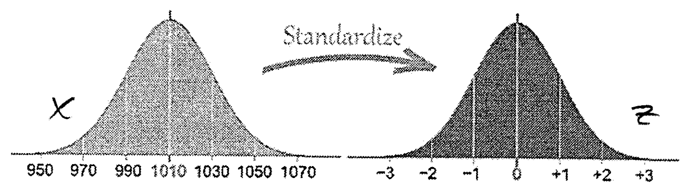
Only Right = $\frac{32\%}{2}$ = $\boxed{16\%}$

8.4 – The Normal Distribution

Finding probabilities based on the normal distribution

- Step 1 → **Standardize** using the **z-score.**

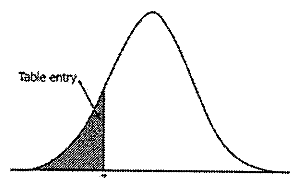
    ☆ Formula: $z = \frac{x-\mu}{\sigma} = \frac{obs - mean}{st\ dev}$ ☆



   - Ex) X has a normal distribution with mean 10 and st dev 2.
     Find the z-score for $X = 13$.

     $z = \frac{13-10}{2} = \boxed{1.5}$

- Step 2 → **Draw, label** and **shade** curve.
    - This is how you show your work!!! ☆



- Step 3 → Use **'Standard Normal Distribution'** table to find the probability for Z.
    - Table ALWAYS gives probability LESS THAN Z: P(Z < z).

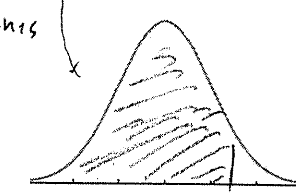
    - Examples → How to use Z-table

matches

    - Left probability = Table (directly)

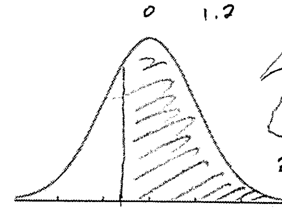
      $P(Z < 1.2) = \boxed{0.8849}$

      1.20



    - Right probability = 1 − Left

      $P(Z > -0.65) =$ 1 − P(Z < −0.65) = 1 − 0.2578
      = $\boxed{0.7422}$



Does NOT match ⟹ extra step

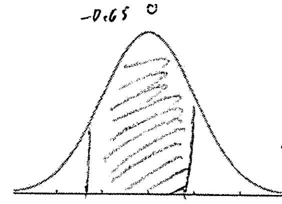
    - Between probability = Left Z_2 − Left Z_1

      $P(-1.13 < Z < 0.75) =$ P(Z < 0.75) − P(Z < −1.13)

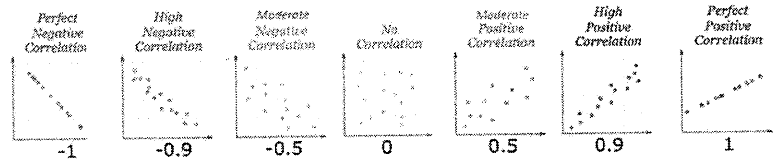
      Z_1 Z_2

      = 0.7734 − 0.1292

      = $\boxed{0.6442}$



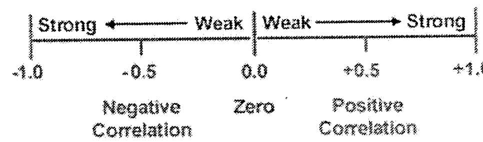
8.5 – Linear Regression



Scatterplots:

- **Form**: Linear, lurved, or random scatter
- **Direction**: Positive, negative or no association
- **Strength**: Weak, moderate or strong

Correlation (r):



- Interpreting correlation (LINEAR)
  - Sign = Direction
  - Absolute value $|r|$ = Strength

- Calculate using calculator
  - **LinReg(ax+b) or 2-Var Stats**
  - $L_1 = X$, $L_2 = Y$

show work by writing this!!

Regression:

- Step 1 → Determine if there is a **significant correlation (linear relationship)**.
  - Compare $|r|$ and Critical Value (CV) for n (sample size) and significance level α.

  ★ If $|r| >$ CV → statistically significant. ★

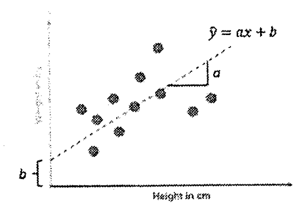
Critical Values of the Pearson Correlation Coefficient

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875

- Step 2 → Once we have a significant correlation, we can find the **regression line**.

  ★ $\hat{y} = ax + b$ (get results from correlation calculation)
  $= slope \cdot x + intercept$ ★

- Step 3 → Make **predictions** using the regression line.
  - Just plug in the new X value to our equation and this will give us the predicted Y.



Example

Dataset:

X	3	5	4	7	6	10
Y	24	40	34	32	17	18

a) Calculate the correlation r.

2-var stats (X, Y)

$\boxed{r = 0.4205}$

OR linreg(ax+b) → X = L_1, Y = L_2

b) Determine if r is significant for $\alpha = 0.01$.

$n = 6$

$|r| = 0.4205 < 0.917 = CV$

⇒ not significant

c) Suppose we have different regression equation where $\hat{y} = 5x + 2$.

Predict Y for $X = 3$:

$\hat{y} = 5(3) + 2 = 17$